International Journal of Speech Technology (2024) 27:739–751
https://doi.org/10.1007/s10772-024-10134-4

Voice pathology detection on spontaneous speech data using deep learning models

Sahar Farazi[1] · Yasser Shekofteh[1]

Received: 3 April 2024 / Accepted: 28 July 2024 / Published online: 10 August 2024

Abstract
Speech problems are a common issue that affects people everywhere and can affect the quality of their lives. The human speech production system involves various components. Dysfunction of any of these components can disrupt normal speech production, giving rise to speech diseases like laryngopharyngeal reflux, vocal cord paralysis, and vocal fold nodules. Early diagnosis of these disorders is very important for the patient's health. Many studies in automatic diagnosis of voice pathology have used sustained vowel sounds and read-speech as the primary source of speech data. However, it is crucial to recognize the unique value of spontaneous-speech. In addition to inheriting the characteristics of read speech, spontaneous-speech offers a more authentic glimpse into individuals' speech behavior. It captures not only linguistic features, but also subtle nuances of human emotions, such as fatigue and excitement, which may cause speech impairments, and shows their patterns in the speech signal better than in the read-speech data. Therefore, we aim to explore spontaneous speech in voice pathology detection to determine if it can help us better understand speech problems. In this research, we examine different deep learning (DL) models trained on two main features (MFCC and Mel spectrograms) for binary classification of healthy speech versus pathological speech, with a specific focus on the spontaneous speech. Extensive experimentation reveals the superiority of our proposed convolutional neural network (CNN) model trained on MFCC features. Notably, the CNN model achieves the highest accuracy, approximately 85% for test data and 92% for evaluation data. These results emphasize the potential of DL approaches in the accurate diagnosis of speech disorders through the analysis of the spontaneous-speech, offering promise for early detection and improved patient care.

Keywords Automatic voice pathology detection · Spontaneous speech · Deep learning · MFCC · Mel-spectrogram · CNN

1 Introduction

Speech-related disorders affect a significant number of individuals in today's world and can impair their quality of life. These disorders are often caused by abnormalities in the human speech production system (SPS) that produces voice and speech (Kent, 2004). These conditions can cause chronic and distressing speech problems, highlighting the need for early diagnosis and intervention. However, many factors, such as time, cost, and limited access to medical equipment can hinder the assessment and treatment process. Moreover, some diagnostic methods such as laryngoscopy can be invasive and uncomfortable for patients. Therefore, there is an urgent demand for the development of non-invasive and widely accessible systems that can accurately detect voice pathology disorders (VPDs) (Shekofteh & Almasganj, 2013, Hegde et al., 2019; Islam et al., 2020, Abdulmajeed et al., 2022, Zhao et al., 2024).

Rapid advances in artificial intelligence (AI) have paved the way for the development and improvement of systems that are capable of automatic and accurate diagnosis of various diseases (Ali et al., 2017, 2018; Chugh et al., 2021; Deepa & Khilar, 2022; Latif et al., 2020). Consequently, there is an urgent need for robust AI-based systems that can perform such diagnoses for VPDs. When an individual experiences abnormalities in their SPS, these changes appear throughout the speech chain, affecting the components involved and altering the characteristics of the speech signal. By extracting relevant speech features from the patient's speech signal, powerful AI-based models can

✉ Yasser Shekofteh
y_shekofteh@sbu.ac.ir

[1] Intelligent Sound Processing Laboratory (ISP-Lab), Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran

identify patterns that distinguish abnormal speech from healthy speech (Abdul & Al-Talabani, 2022; Abdulmajeed et al., 2023; Shekofteh & Almasganj, 2013). In this research, we use *Mel*-Frequency Cepstral Coefficients (MFCCs) and Mel-spectrograms as input features extracted from the spontaneous-speech signals and use deep learning (DL) models to classify people as healthy or with voice disorders.

While many studies in voice pathology detection tasks mainly use sustained vowel speech data, which facilitates easier comparison of results, it is worth noting that some researchers had also successfully used the read-speech data and achieved high accuracy in the VPDs tasks (Ali et al., 2013). This choice is rooted in the belief that continuous speech serves as a reliable indicator of normal speech, closely mirroring everyday conversational habits. The read-speech encompasses a wide spectrum of speech elements, including vowels, consonants, and their combinations, making it a valuable resource for extracting features pertinent to both healthy and pathological speech patterns.

In contrast, we chose to focus on the spontaneous-speech data for our study. We hold the belief that the spontaneous-speech represents the most natural form of everyday communication, encompassing combinations of vowels, consonants, and moments of silence, while also reflecting the emotional nuances experienced during the speech. Unlike the read-speech, the spontaneous-speech can reveal certain pathological features that some patients may be able to mask when reading a pre-prepared text. But they are not able to hide and control them during the spontaneous speech. By examining the spontaneous-speech data, researchers and clinicians may gain a more holistic perspective on an individual's communication abilities, allowing for the detection and characterization of a broader range of voice disorders and related issues in a real-world context.

To the best of our knowledge, no previous studies have investigated the use of the spontaneous-speech data to detect voice abnormalities and distinguish between healthy individuals and individuals with voice disorders. Most studies have focused primarily on vowel data. One of the reasons for this emphasis on vowel data is the ease of comparing results with previous studies, since vowel production exhibits minimal variation across languages, dialects, and accents across individuals and languages. However, some studies have incorporated continuous speech data, such as the read-speech, into their research and found that this type of data yields improved results. Continuous speech is more representative of real-life situations because people use it more during their daily conversations compared to producing isolated vowels (Ali et al., 2013). Furthermore, combining speech sounds, including vowels, consonants, and pauses, and extracting features from these types of data, although more challenging, can potentially provide more detailed insight into a patient's speech and voice states. Moreover, we believe that the spontaneous-speech can provide clearer indications of speech abnormalities and disorders.

In the spontaneous-speech, unlike the read-speech, individuals do not read or recite predetermined sentences, so they cannot change their speaking style or control their voice problems. Instead, they speak spontaneously, which is the most natural form of human speech. While there has been limited research on the spontaneous-speech data in this field, we include this type of data in our research. We aim to evaluate the effect of the spontaneous-speech on the automatic voice pathology detection task. Therefore, our main goal in the present study is to use the spontaneous-speech data to see how it affects classification accuracy in VPDs tasks using DL models.

The DL models have a rich history that has seen significant advances in various applications, including their profound impact on health data analytics, particularly in the domain of automatic voice pathology detection. They have enabled the development of sophisticated algorithms that can analyze subtle features in vocal patterns to identify and classify voice disorders with remarkable accuracy. Notable DL models frequently employed in voice pathology detection and classification include deep forward neural networks, and convolutional neural networks (CNNs) which are excellent at capturing spatial dependencies in audio data and its two-dimension-extracted features such as *Mel*-spectrograms and MFCCs of the audio frames, and recurrent neural networks (RNNs), which are renowned for their ability to model sequential patterns in the speech. Consequently, our main goal is to increase classification accuracy and improve our understanding of the effectiveness of spontaneous-speech data. We aim to identify the DL models and specific features that can collaboratively better the detection of voice pathology.

Our proposed model for classifying healthy subjects and those with pathological speech is shown in Fig. 1. It utilizes raw waveforms of healthy and impaired speech signals. After the preprocessing steps, which are explained in detail later in the paper, we extract MFCC and *Mel*-spectrograms from each frame. Subsequently, we apply three kinds of DL models to distinguish healthy samples from those with voice disorders.

In the field of automatic diagnosis of voice pathology, limited and free databases are available. The *Massachusetts Eye and Ear Infirmary Database (MEEI)* in English, the *Saarbruecken Voice Database (SVD)* in German, and the *Arabic Voice Pathology database (AVPD)* are among the most well-known databases, which have been frequently utilized in related studies (Hegde et al., 2019; Lee, 2021; Mesallam et al., 2017; Sindhu & Sainin, 2024). These databases mainly consist of audio files containing vowel sounds and speech including reading sentences or texts. In our study, we employed the dataset of *Advanced Voice Function*

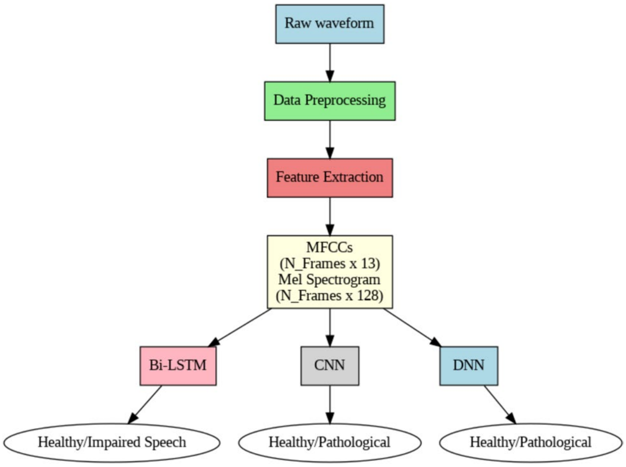

Fig. 1 The proposed classification flowchart of healthy people and patients in this study

Assessment Databases (AVFAD) in Portugal, collected by the *University of Aveiro*, which includes voice recordings from people with various speech disorders as well as healthy voice individuals (Jesus et al., 2017). In addition to sustaining vowels and read-speech data, the *AVFAD* also has spontaneous-speech data for each sample.

The novelties of the proposed method in this paper are outlined below:

(a) Utilizing the spontaneous-speech data as the primary data type for automatic voice pathology detection.
(b) Employing the *AVFAD* dataset, which comprises diverse pathology types (approximately 26), for both training and testing of our deep learning models, and identifying the most effective model.

The remainder of this paper is organized as follows: we will first review the related works in Sect. 2. Then, we introduce the data used, feature extraction methods, and the proposed DL models employed in this study. Subsequently, we present evaluations and results, and finally, we discuss the conclusions of our study and outline future works.

2 Related works

Numerous studies have been conducted in the field of voice pathology and speech disorder diagnosis with the aim of distinguishing between normal and abnormal voices. These studies primarily focus on the binary classification and distinguishing between healthy individuals and individuals with voice disorders. In addition, some studies make multiple classifications and categorize individuals according to the specific type of voice disorder they have such as vocal nodules, vocal polyps, laryngitis, and vocal cord paralysis (Payten et al., 2022).

In automatic VPDs detection, researchers can utilize three types of speech data: sustained vowel data, read-speech data, and spontaneous-speech data. Sustained vowel data refers to recordings or samples of individuals producing and sustaining specific vowel sounds for analysis. These sustained vowels are often chosen from a set that includes commonly used vowel sounds in human speech, such as /a/, /e/, /i/, /o/, and /u/. The *SVD*, *AVPD*, *MEEI*, and *AVFAD* datasets contain this type of data.

The purpose of collecting sustained vowel data in voice pathology detection is to assess and analyze various acoustic properties and characteristics of the sustained vowels. Changes in these acoustic features can provide valuable information about potential voice disorders or abnormalities. For instance, in (Mohammed et al., 2020), the *SVD* dataset was utilized, focusing solely on the vowel /a/. They employed a pre-trained CNN, specifically the ResNet34 model, with a spectrogram as its input. Their spectrogram consisted of at least 20 band filters based on octaves. They achieved about 96% accuracy in automatic VPDs detection by their system. In (Ksibi et al., 2023), they used the SVD dataset, focusing on the vowel /a/. The aim of this study was to create an automatic voice pathology detection system for early detection of voice abnormalities and specific pathologies. For binary classification of healthy versus unhealthy samples, manual audio features such as MFCC, ZCR, and RMSE were extracted. A two-level classifier model was employed, where the first level classifies male samples from female samples and the next level distinguishes healthy samples from unhealthy ones. Their model consisted of CNN blocks followed by an RNN block, with two dense layers after the RNN. The proposed model achieved an accuracy of 88.84%. In (Verma et al., 2023), the vowel /a/ sound from the VOICED dataset was used. This research proposed a novel methodology called VDDMFS (Voice Disorder Detection using MFCC, Fundamental frequency, and Spectral centroid). It combined an artificial neural network (ANN) trained on acoustic attributes and a long short-term memory (LSTM) model trained on MFCC attributes. The probabilities generated by both the ANN and LSTM models were then stacked and used as input to XGBoost, which detected whether a voice was disordered or not. This approach achieved an accuracy of 95.67%. In (Muhammad & Alhussein, 2021), researchers utilized the *SVD* dataset, specifically focusing on data associated with sustained vowel sounds, particularly the vowel sound /a/. In addition to the signal related to the vowel sound /a/, they incorporated the Electroglottogram (EGG) signal extracted from the subjects. The input data for their model comprised spectrograms and *Mel*-spectrograms extracted from both the vowel sound signal and the corresponding EGG signal. Their model was composed of two pre-trained CNNs: one takes as input the spectrograms extracted from the vowel sound signal, while the other CNN's input was the spectrogram extracted from the EGG signal. The outputs of the CNNs were then combined and fed into a bidirectional long short-term memory (Bi-LSTM) network for classification. They achieved their highest accuracy using the *Xception* model as their pre-trained CNN model. They were able to attain a classification accuracy of 95.65% for distinguishing between healthy and pathological subjects.

In voice pathology detection, read-speech data refers to recordings or samples of individuals reading a specified text or passage aloud. This type of speech data is collected to analyze a person's voice during natural and conversational speech patterns. The *SVD*, *AVPD*, *MEEI*, and AVFAD datasets have this type of data. The read-speech is more natural and reflects the way individuals typically communicate (Hegde et al., 2019). For example, in (Ali et al., 2013) the *AVPD* dataset was used, specifically the read-speech data. In this research, the MFCCs were employed as features and inputted into a Gaussian Mixture Model (GMM), resulting in a notable accuracy of 91.66%. In (Ali et al., 2016), the study employed the read-speech data from the *MEEI* database to classify healthy subjects from pathological subjects. Two types of features were utilized in this research: auditory spectrum features and All-Pole model-based cepstral coefficients (APCC). This study achieved a notable accuracy of 99.56% for binary classification, employing 36 APCC coefficients (12 static and its first and second derivatives) along with an 80-GMM for classification. In Narendra and Alku (2020), researchers compared traditional pipeline and end-to-end approaches using glottal source information to investigate automatic methods for detecting pathological voice from healthy speech. They used both sustained vowel and read speech data types. In the traditional pipeline approach, two sets of glottal features and openSMILE features were used to train support vector machine (SVM) classifiers. The end-to-end approach utilized raw speech signals and glottal flow waveforms to train two DL architectures: a combination of CNN and multilayer perceptron (MLP), and a combination of CNN and LSTM network. The end-to-end approach showed a 2–3% improvement in accuracy when using glottal flow as input compared to the raw speech waveform.

The spontaneous-speech data in voice pathology detection refers to recordings or samples of individuals engaging in unscripted, natural conversations or monologues. Unlike controlled tasks such as sustained vowel production or reading a text aloud, the spontaneous-speech involves the free-flowing expression of thoughts and ideas in real-life communication scenarios. Analyzing of the spontaneous-speech data can be valuable in voice pathology detection for several reasons. This type of data is important for understanding how voice disorders may impact communication in diverse and unstructured contexts. In addition, the spontaneous-speech as a continuous speech contains prosodic features, including intonation, rhythm, and stress patterns. Changes in prosody can be indicative of certain voice disorders or emotional states that affect speech. Also, the spontaneous-speech data contains different kinds of vowels, consonants, silence, and their combinations, and it is another positive point of this type of data for detecting speech disorders. There are some works that have used AI models and spontaneous speech data to distinguish healthy speech from the impaired speech caused by a specific disease such as Alzheimer's or Parkinson’s. For example, in a study by (Chen et al., 2021), they used spontaneous speech data and employed a logistic regression model with acoustic and linguistic features to classify cognitively healthy individuals and patients with Alzheimer’s disease (AD). However, we have not found any study that addresses the gap of using spontaneous speech data to classify healthy and unhealthy samples containing various impairments related to different diseases. One of our initial focuses in this research is to develop a model with the ability to learn and notice patterns in speech signals that may indicate impairments caused by different types of diseases. Therefore, in this study, we investigated the impact of incorporating spontaneous-speech data to distinguish pathological samples from healthy samples.

The number of reported studies using the *AVFAD* database for automatic voice pathology detection and classification is very small compared to other well-known databases, and according to our studies, the spontaneous speech has not been evaluated in any of them. In (Oliveira et al., 2020), researchers employed the *AVFAD* and *SVD* databases to investigate the effect of the combination of vowel sounds (/a/, /i/, and /u/) on extracted wavelet coefficients. They utilized a random forest (RF) machine learning model for classification and found that combining vowel sounds enhanced the distinguishability of wavelet coefficients extracted from the audio signals, resulting in a more accurate classification of healthy and abnormal voices. Therefore, combining vowel sounds was beneficial in increasing the discriminative properties of the extracted wavelet coefficients, which ultimately led to improved accuracy in the classification of healthy and unhealthy samples. In (Ribas et al., 2023a), the *SVD* and *AVFAD* datasets were used. Their data was prepared from the type of sustained vowel /a/ and the reading-speech from the reading of some sentences. In this study, they used an SVM learning model to classify pathologic samples from healthy ones. Their extracted features included several common features in this field such as MFCCs. Their best accuracy using the read-speech part of the SVD dataset was around 95%, and their best accuracy was around 91.5% on the *AVFAD* dataset, which was achieved with the sustained vowel. In (Ribas et al., 2023b, b), the *SVD*, and *AVFAD* datasets were utilized to train the model. The primary objective

Table 1 Data division methodology of the AVFAD dataset

Data	Train		Test		Validation	
Gender	Male	Female	Male	Female	Male	Female
Normal	73	162	22	50	18	37
Pathologic	64	161	20	49	13	37

of this research was to encompass various disorders within their patient samples. However, due to the limited number of data instances for certain types of diseases in the *SVD*, the researchers augmented their dataset by incorporating the *AVFAD* dataset. The results of the study demonstrated a significant improvement in classification accuracy with the addition of *AVFAD*. To extract features from audio signals, the researchers utilized self-supervised pre-trained models, namely wav2vec2, HuBERT, and WavLM. Subsequently, they applied once a transformer and once a DNN for classification. Notably, the highest accuracy was achieved when HuBERT was used for feature extraction, paired with a transformer for classification. They achieved a remarkable accuracy of 94.27% for the classification of pathological and healthy samples.

3 Materials and methods

In this section, the data set used in this research is explained. It also provides detailed information on data preparation steps, preprocessing, and feature extraction methods, and proposed architectures for DL models used in classification.

3.1 Study subjects

In our study, we employed the *AVFAD* dataset, which was collected by the *University of Aveiro* in Portugal and consists of recorded voices from individuals with various speech disorders as well as individuals with healthy voices. The *AVFAD* dataset encompasses three types of data for each person. The first type is related to sustained vowel sounds, especially the vowel sounds /a/, /u/, and /i/. Each audio file for each sustained vowel sound contains three repetitions of that vowel sound.

The second type focuses on reading a specific text or sentence, comprising seven parts that involve reading six *Portuguese CAPE-V* sentences (Association, 2009; Jesus et al., 2009). Each sentence from the read-speech part of the *AVFAD* dataset was read three times by each subject; hence, each sentence audio file contained three repetitions of that sentence. The seventh part was related to reading a famous phonetically balanced text which was the Portuguese version of the passage "*The North Wind and the Son*" (Jesus et al., 2009). In our investigation, we utilized data from the spontaneous-speech data. The sampling frequency of all audio files in this dataset remained at 48 kHz.

The dataset consists of a total of 709 individuals, with 346 clinically diagnosed with vocal pathology and 363 without any vocal alterations. We used data from 706 samples because we encountered problems with the files of three samples and decided to remove them from the data set entirely.[1] For our study, we partitioned the data into training, test, and evaluation sets with ratios of 65%, 20%, and 15%, respectively (*Appendix 1*). The aim of this partitioning is to achieve the optimal balance among these sets regarding the gender distribution and the proportion of impaired and healthy people. The distribution of data based on gender and health status is presented in Table 1. In total, 460 individuals' audio files were utilized for training, 105 for validation, and 141 for testing purposes.

The minimum time interval for the spontaneous-speech data files is approximately 20 s, while their maximum duration extends to about 338 s, roughly equivalent to 5 min. On average, these files span about 54 s, or about one minute.

3.2 Feature extraction

In this study, we extracted the first 13 coefficients of MFCCs as input features from the voice signals of the samples. MFCCs are widely recognized as essential features in the field of voice pathology detection. To obtain the MFCCs from the audio signals, we employed a series of signal processing steps. These steps involved applying a pre-emphasized filter, windowing the signal, performing a *fast Fourier transformation*, applying a *Mel* filter bank, and subsequently applying nonlinear step and Discrete Cosine transformations (DCT). We utilized the Librosa library in Python to compute the MFCC features.[2] Each voice frame had a size of 42 ms, with a hop length of 512 samples. Furthermore, the sampling frequency of each audio file was set to 48,000 Hz. Mathematically, the MFCC computation using a filterbank with M filters can be represented as follows:

$$MFCC[i] = \sum\nolimits_{j=0}^{M-1} log\left(\sum_{k=0}^{N-1} |X[k]|^2 H_m[k]\right)\cos(\frac{\pi i(j+0.5)}{N}) \quad (1)$$

[1] Three files' name that were excluded: MBC, MEX,MLF.

[2] Available at: https://librosa.org/.

where *MFCC[i]* is the *i*-th coefficient, $X[k]$ is the k-th bin of Fourier transform of the signal, N is the number of FFT points, $H_m[k]$ is the *Mel* filterbank, and cos(.) denotes the cosine function.

We have also used *Mel*-Spectrograms from the wave audio files as another type of input feature for our proposed DL models. We examined which input feature worked better for our goal. Initially, we divided the audio signals into short overlapping frames, typically using a frame size of 42 ms with an approximately 10-ms overlap. For each frame, we calculated the *Short-Time Fourier Transform (STFT)* to obtain the magnitude spectrum. From there, we applied a *Mel* filter bank with 128 frequency bins, capturing the relevant spectral information across a wide frequency range. These *Mel*-spectrograms serve as a compact informative representation of the speech signals, preserving its essential spectral characteristics. We utilized the Librosa library in Python to extract Mel-spectrograms.

Given that the duration of the spontaneous-speech audio files varies across different individuals in the dataset, we adopted a strategy to address this variability. We calculated the value of S, which represents the sum of the value of average number of frames and the value of standard deviation of the number of frames for all files.

$$S = Average number of frames + Standard deviation of the number of frames$$

Therefore, in this study, the first 83 s of the spontaneous-speech of each individual are used for our models. Consequently, the number of frames for which we extracted coefficients might be less than the total number of frames in some files. Conversely, for certain files, the value of S exceeded the original number of frames. In such cases, we padded each frame with zeros to compensate for the missing values.

All input data are normalized, a crucial preprocessing step that ensures uniformity in the data's scale. Normalization significantly enhances the training process by allowing the model to learn more effectively from the features, contributing to improved overall performance. Therefore, in this study, to increase the learning of the models, the data are normalized using the z-score normalization method.

In z-score normalization, values are adjusted relative to the mean and standard deviation of attribute A. For a given value V_i of attribute A, the normalized value U_i is calculated as follows:

$$U_i = \frac{V_i - Avg(A)}{Std(A)} \tag{2}$$

where V_i is the original value of attribute A, *Avg* (A) represents the average (mean) of all values of attribute A, *Std* (A) denotes the standard deviation of all values of attribute A.

As a result, our data were formed as two-dimensional (2D) matrices ($S \times 13$) for MFCCs and ($S \times 128$) for *Mel*-Spectrograms, where 13 is the number of MFCCs and 128 is the number of Mel frequency bands. Figure 2 shows the *Mel*-Spectrogram of a healthy and a pathologic sample's spontaneous-speech signal from the *AVFAD* dataset, and Fig. 3, shows the MFCCs heatmap of a healthy and a pathologic sample's spontaneous-speech signal From the *AVFAD* dataset.

3.3 Deep learning models

In this study, we employed a Convolutional Neural Network (CNN), a Bidirectional Long Short Term Memory (Bi-LSTM) model, and a Deep Neural Network (DNN) for learning data and classification (Bengio et al., 2017; Zhang et al., 2023). The structure and parameters of these models have been selected after many tests on our dataset to get the best results. In all three models, to prevent overfitting, we applied dropout layers (Srivastava et al., 2014). Also, we employed binary cross-entropy as the loss function. This loss function is commonly used for binary classification tasks, facilitating the differentiation between healthy and unhealthy samples with greater accuracy. We also employed the popular *Adam* optimizer for the training process.

3.3.1 Convolutional neural network (CNN)

CNNs have shown remarkable performance in various domains, including image and speech processing (Gu et al., 2018; Li et al., 2021; Mohammed et al., 2020). They excel at capturing the spatial hierarchies of features in the data, which makes them well-suited for analyzing our 2D-MFCC features and *Mel*-Spectrograms. The convolution layers in the CNNs are able to automatically learn the relevant patterns in the spectrogram and MFCCs, helping to identify distinctive features associated with voice pathology. We utilized three convolutional layers along with two 2D-maxpooling layers for the feature extraction. A 2D-global average pooling layer was employed to aggregate the extracted features. Based on our evaluations, using the *Global Average Pooling* layer instead of the Flatten layer in the convolutional neural network using the data of this study has improved the performance of final networks (Al-Sabaawi et al., 2020). Subsequently, two dense layers were implemented, with the last one incorporating a sigmoid activation function for the final classification. The proposed model is implemented using the Keras library and TensorFlow framework in Python. Our proposed CNN

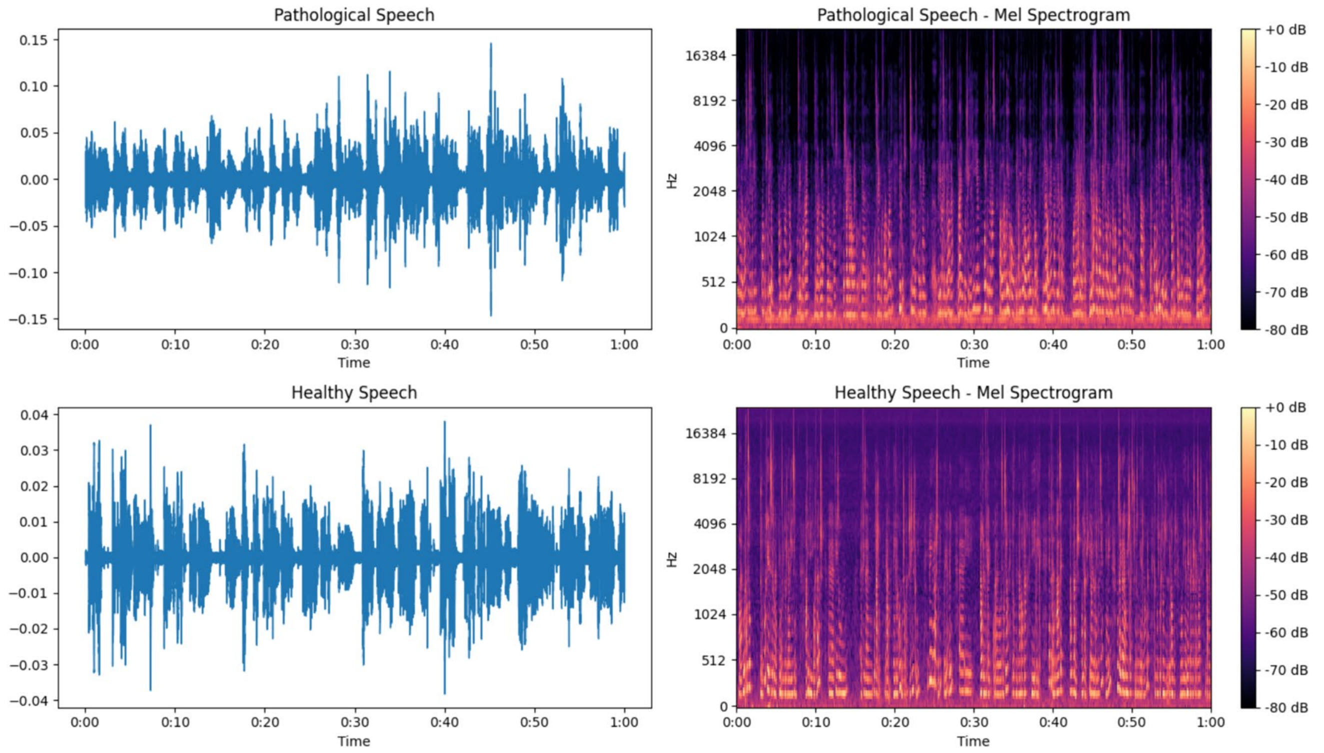

Fig. 2 *Mel*-spectrogram and waveform comparison of the spontaneous-speech data for a pathological individual (up) and a healthy individual (bottom) for their first 60 s

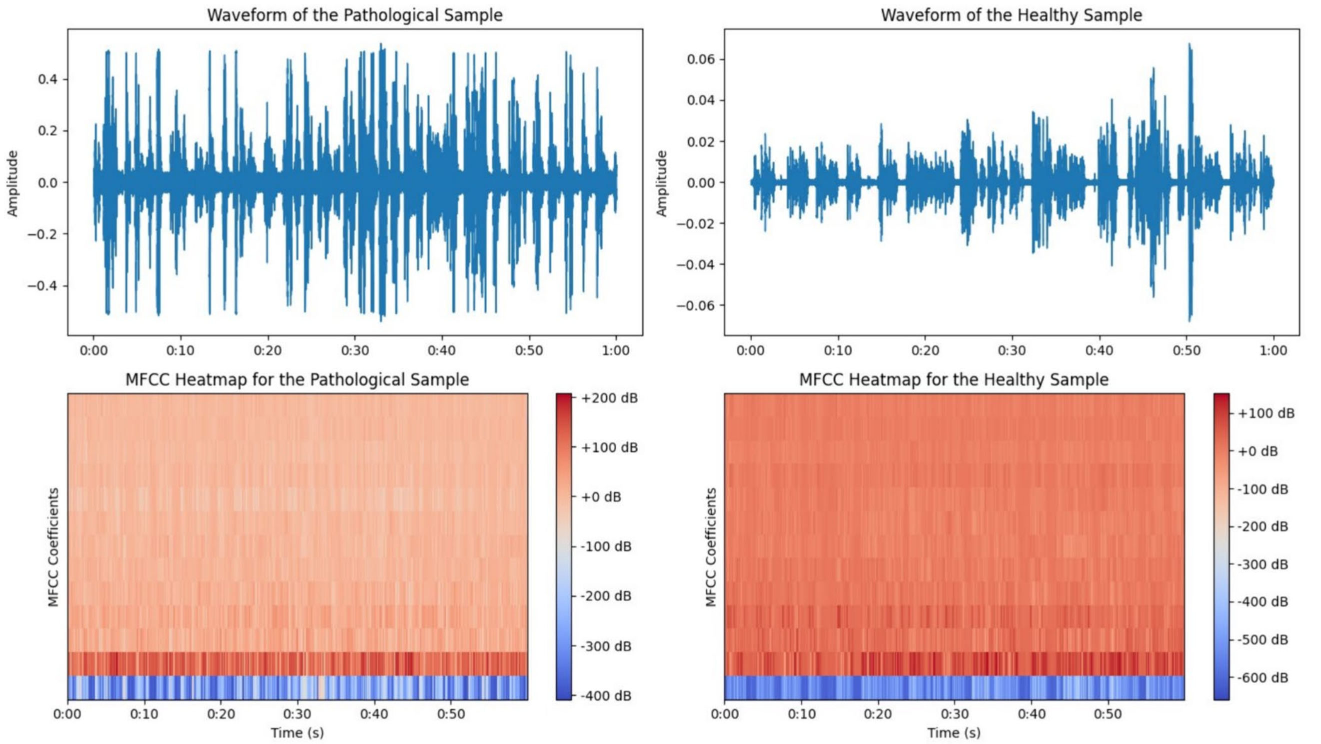

Fig. 3 MFCCs heatmap and waveform comparison of the spontaneous-speech data for a pathological individual (right) and a healthy individual (left) for their first 60 s

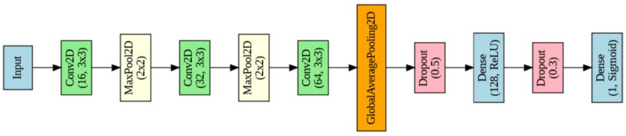

Fig. 4 The proposed CNN Model for VPD detection

model is shown in Fig. 4, and its details are explained in Table 2.

3.3.2 Bidirectional long short-term memory (Bi-LSTM)

The bidirectional LSTM (Bi-LSTM) architecture is selected for its ability to effectively capture temporal dependencies in sequential data (Graves & Schmidhuber, 2005; Graves et al., 2005; Syed et al., 2021), a quality that proves highly beneficial when working with the MFCCs. This is particularly relevant as the MFCCs, like the *Mel*-spectrograms, represent a type of sequential data, encapsulating the coefficients of each frame within a speech signal. Our proposed LSTM model consists of a single layer of bidirectional LSTM with 128 neurons, accompanied by a dropout layer. Additionally, the architecture incorporates two dense layers: the first with 32 neurons and the second with 1 neuron, serving the specific purpose of the binary classification. The proposed model was implemented using the Keras library and TensorFlow framework in Python. Figure 5 shows the proposed Bi-LSTM model, and its details are explained in Table 3.

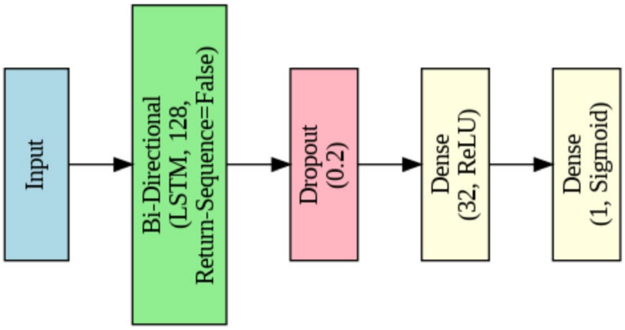

Fig. 5 The proposed Bi-LSTM Model for VPD detection

3.3.3 Deep neural network (DNN)

A DNN model with multiple dense layers can be advantageous for voice pathology detection (Ankışhan & İnam, 2021; Chen & Chen, 2022; Chuang et al., 2018; Zakariah et al., 2022). While it may not have the same level of sequence modeling capabilities as RNNs like Bi-LSTM, the DNNs can still learn complex relationships within high-dimensional feature vectors derived from the MFCCs or *Mel*-Spectrograms. By utilizing multiple dense layers, the DNN can capture intricate patterns and non-linear relationships in the data. Our DNN model starts with a flattened layer to input the model, followed by five dense layers. We have also added a dropout layer to improve the model's robustness. The proposed model was implemented using the Keras library and TensorFlow framework in Python. The proposed DNN is shown in Fig. 6, and its details are explained in Table 4.

4 Evaluation and results

To ascertain the robustness and reliability of our experimental findings, we developed a rigorous validation protocol. Each model was subjected to three independent executions on our dataset. For each run, the models processed the data and performance metrics were recorded. After completion of the runs, we aggregated the results to calculate the mean performance metrics, expressed as average values ± standard deviation. This statistical approach

Table 2 The proposed CNN model's parameters and details

Input Layer	Input shape: (n_frames, 13 or 128, 1)
Convolutional Layer 1	Kernel size: 3 × 3, Filters: 16, Activation: ReLU, Padding: same
MaxPooling Layer 1	Pool size: 2 × 2
Convolutional Layer 2	Kernel size: 3 × 3, Filters: 32, Activation: ReLU, Padding: same
MaxPooling Layer 2	Pool size: 2 × 2
Convolutional Layer 3	Kernel size: 3 × 3, Filters: 64, Activation: ReLU, Padding: same
2D Global Average Pooling	
Dropout Layer 1	Dropout rate: 0.5
Dense Layer 1	Units: 128, Activation: ReLU
Dropout Layer 2	Dropout rate: 0.3
Dense Layer 2	Units: 1, Activation: Sigmoid
Optimizer: Adam, Learning rate: 0.001	
Loss Function: Binary Cross-Entropy	

Table 3 The proposed Bi-LSTM Model's parameters and details

Input Layer	Input shape: (n_frames, 13 or 128)
Bidirectional LSTM Layer	Units: 128, Return sequences: False
Dropout layer	Dropout rate: 0.2
Dense layer	Units: 32, Activation: ReLU
Dense layer	Units: 1, Activation: Sigmoid
Optimizer: Adam, Learning rate: 0.001	
Loss function: Binary Cross-Entropy	

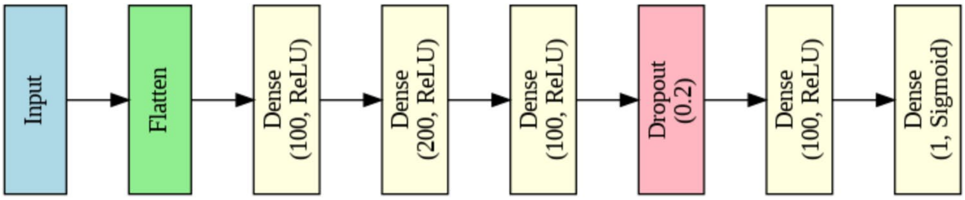

Fig. 6 The proposed DNN model for VPD detection

Table 4 The proposed DNN Model's parameters and details

Input layer	Input shape: (n_frames, 13 or 128)
Flatten layer	–
Dense layer	Units: 100, Activation: ReLU
Dense Layer	Units: 200, Activation: ReLU
Dense layer	Units: 100, Activation: ReLU
Dropout layer	Dropout rate: 0.2
Dense layer	Units: 100, Activation: ReLU
Dense layer	Units: 1, Activation: Sigmoid
Optimizer:	Adam, Learning rate: 0.001
Loss function: binary cross-entropy	

Table 5 Classification results of three proposed DL-based models using the *Mel*-spectrogram and MFCC features for the spontaneous-speech data

Feature	Model	Accuracy of validation data (mean ± std)	Accuracy of test data (mean ± std)
MFCCs	CNN	**0.9206 ± 0.0119**	**0.8510 ± 0.0100**
Mel-Spectrogram	CNN	0.8578 ± 0.0005	0.8301 ± 0.0266
MFCCs	Bi-LSTM	0.8793 ± 0.0045	0.7872 ± 0.0116
Mel-Spectrogram	Bi-LSTM	0.7364 ± 0.0273	0.6926 ± 0.0261
MFCCs	DNN	0.8444 ± 0.0119	0.7399 ± 0.0203
Mel-Spectrogram	DNN	0.7142 ± 0.0155	0.6500 ± 0.0066

The best evaluation values are in bold

provides a measure of central tendency and variability and provides insights into the consistency of the model's performance across multiple experiments.

The adoption of this methodology is pivotal in mitigating the influence of outliers and anomalies, thereby yielding more stable and representative metrics. It also allows for a detailed evaluation of the models' performance, particularly in their ability to process and analyze two distinct types of acoustic features: the MFCCs and *Mel*-Spectrograms. By conducting a comparative analysis of the models' performance on these two feature types, we can draw comprehensive conclusions about their respective strengths and limitations. Table 5 shows the accuracy metric obtained from our models on two types of input data for the DL-based models.

The outcomes of our experiments demonstrate that the proposed CNN is superior to the other two network architectures we investigated. Across both the evaluation and test datasets, the CNN consistently outperforms these models showing its strength and effectiveness in our

Table 6 Performance metrics of the Best-Performing run of CNN model with MFCCs for healthy and pathological speech classes on the test dataset for the spontaneous-speech data

Label	Precision	Recall	F1-Score
Healthy speech	0.91	0.82	0.86
Pathological speech	0.83	0.91	0.87

study. Moreover, our analysis of feature representations has provided insights. When comparing the effect of the *Mel*-Spectrogram and MFCCs as the models' features, it is clear that the model trained on the MFCCs excels in classifying healthy and unhealthy samples.

Among the three runs of our CNN on the MFCCs, the best accuracy achieved for the test dataset is approximately 87%. Table 6 shows the performance metrics of the *best-performing runs of CNN model* with MFCCs among three runs on the test dataset for healthy and pathologic samples. The metrics we obtained highlight our model effectiveness, in categorizing individuals as either *'Healthy Speech'* or *'Pathological Speech'*.

The precision value of 0.91 for the *'Healthy Speech'* class demonstrates the model's accuracy in identifying healthy subjects, indicating that 91% of the instances predicted as *'Healthy Speech'* were indeed healthy. Similarly, the model exhibited a precision of 0.83 for the *'Pathologic'* class, signifying its ability to correctly identify 83% of the instances predicted as *'Pathological Speech'*.

Regarding recall, with a value of 0.82 for the *'Healthy Speech'* class, the model successfully identified 82% of all actual healthy instances out of the total healthy instances in the test dataset. For the *'Pathological Speech'* class, the recall value of 0.91 indicates that the model accurately detected 91% of all actual pathological instances out of the total pathological instances in the test dataset, indicating its ability to comprehensively identify positive cases.

The F1-score, which balances precision and recall, provides a consolidated measure of the model's overall performance. Based on the results of Table 3, the F1-score of 0.86 for the *'Healthy Speech'* class and 0.87 for the *'Pathological Speech'* class signifies a robust balance between precision and recall, indicating a high level of accuracy and reliability in the models' predictions for the spontaneous-speech data.

In this experiment, the best-performing run of the CNN model on the MFCCs achieved an accuracy of approximately 87%. However, six genuinely pathological cases in the test dataset were misclassified as healthy by our model. Table 7 presents the disease names of these misclassified pathological samples, along with the percentage ratio of the frequency of misclassification to the occurrence of these diseases in the test dataset and the occurrence of the corresponding diseases in the training dataset.

As shown in Table 7, the highest misclassification rate occurred for "*Vocal Varices and Ectasia*". This is justified by the extremely low occurrence of this disease in the training dataset, accounting for only 4.4%. Consequently, the model's error rate for identifying this specific condition is reasonably higher than for other diseases. On the other hand, the lowest misclassification rate was for "*Laryngopharyngeal Reflux*". Approximately 30% of the samples in the test dataset were affected by this disease, and the model was trained on a sufficient number of instances representing this condition. Therefore, the low misclassification rate for this disease is justifiable, given sufficient training data available for the model to learn from.

In summary, these misclassification patterns highlight the influence of disease prevalence in both the training and test datasets on our model's performance. Rare diseases in the training dataset can lead to higher misclassification rates, while diseases with adequate representation in the training data result in more accurate predictions. This analysis provides valuable insights into the challenges faced in voice pathology detection and emphasizes the importance of balanced and representative training datasets for robust model performance.

5 Discussion and conclusions

To the best of our knowledge, our study represents a pioneering effort in utilizing the spontaneous-speech data for the task of automatic voice pathology detection and classifying healthy individuals from those potentially experiencing

Table 7 Analysis of pathologic samples misclassified as healthy by the best-performing CNN

Disease	Number of misclassification	Percentage ratio of misclassification repetition to disease occurrences in the test dataset	Percentage of disease occurrences in the training dataset
Vocal Cord Nodules	1	16.7%	8.0%
Vocal Hemorrhage	1	33.3%	2.7%
Laryngopharyngeal Reflux	2	10.5%	30.2%
Vocal Cord Cyst	1	14.3%	5.3%
Vocal Varices and Ectasia	1	50.0%	4.4%

voice disorders. The absence of prior research in this specific domain makes direct comparisons challenging. However, our achieved accuracies provide compelling evidence for the efficacy of utilizing the spontaneous-speech data in this context.

Our results strongly indicated that incorporating the spontaneous-speech as a speech type for analysis provided significant benefits. The use of this most natural type of speech increased the accuracy and robustness of voice pathology diagnosis. The novel approach presented in this study not only contributes to the advancement of the field but also emphasizes the potential of the spontaneous-speech data in improving the accuracy and reliability of voice pathology detection methodologies.

Looking forward, there are promising avenues for future research. Exploring more robust machine learning models and incorporating diverse types of features derived from the speech signals can further enhance the performance of voice pathology classification. By exploring the application of advanced models and using a wider range of features, we anticipate significant improvements in the accuracy and sophistication of voice pathology detection systems.

Appendix 1

Files extracted and their data division from the AVFAD dataset.

Train set files' name (460 files):

Normal samples' file name (235 files)	Pathological samples' file names (225 files)
'ABR' 'AAF' 'ACN' 'ACA' 'AAP' 'AFP' 'AFS' 'AFB' 'AGQ' 'AFF' 'ACV' 'AJO'	'AAM' 'AAC' 'ACM' 'AAY' 'AAX' 'ADM' 'AJS' 'AJB' 'AJX' 'AJC' 'ALX' 'ALF'
'AIA' 'AJP' 'ALK' 'AMF' 'ALY' 'AMN' 'AMW' 'AMX' 'APN' 'APP' 'APH' 'ANA'	'ALR' 'APM' 'AMZ' 'ARX' 'AQC' 'ARP' 'APX' 'APZ' 'ASX' 'ASR' 'ATA' 'AVC'
'ASM' 'ARV' 'ARK' 'ARS' 'ASP' 'ASF' 'AVG' 'BMA' 'CAX' 'AXM' 'CCW' 'CCS'	'CMF' 'DFC' 'DOX' 'DMG' 'DGM' 'ECF' 'EJA' 'FFS' 'FAX' 'FAS' 'FMA' 'FMX'
'CCO' 'CGM' 'CFW' 'CMD' 'CLR' 'CJG' 'COF' 'CMS' 'CMT' 'CSF' 'CRR' 'CSM'	'FMP' 'FMC' 'FOL' 'FRL' 'FRM' 'GFN' 'HFV' 'ICX' 'JAB' 'JBG' 'JCS' 'JAS'
'CSY' 'CSR' 'CSX' 'DCS' 'DCC' 'DRX' 'DMR' 'DGF' 'DFR' 'EAS' 'ELM' 'ECC'	'JCA' 'JAO' 'JAM' 'JAP' 'JDM' 'JLF' 'JGC' 'JJF' 'JMB' 'JMJ' 'JML' 'JMF'
'ERS' 'FCG' 'FCS' 'ESR' 'ESB' 'FFX' 'GHA' 'GCR' 'GJM' 'GRS' 'GJG' 'HCB'	'JOL' 'JMM' 'JMX' 'JMP' 'AAS' 'ACS' 'ABS' 'AFL' 'ACC' 'ACO' 'AAO' 'AFR'
'HCC' 'HCS' 'HGP' 'ICA' 'HMR' 'ICV' 'IDB' 'ICM' 'IFG' 'IMN' 'ISM' 'IJM'	'AGC' 'AIT' 'ALB' 'AGA' 'AFQ' 'AMA' 'AMB' 'AML' 'AMT' 'AMP' 'AMS' 'AMO'
'IMB' 'IGM' 'JCF' 'IVP' 'ISP' 'JFF' 'JCX' 'JLS' 'JMR' 'JSF' 'JTM' 'LCM'	'AMM' 'AMC' 'ALC' 'AMY' 'AMV' 'APD' 'APA' 'APR' 'ASL' 'BJA' 'ASS' 'ASN'
'LCB' 'LCG' 'LFB' 'LCS' 'LJC' 'LML' 'LMM' 'LMS' 'MAB' 'LSP' 'MAJ' 'MAN'	'CBT' 'BPB' 'CCR' 'CAS' 'CAC' 'CAG' 'BPS' 'BVM' 'CFS' 'CAN' 'CMG' 'DCV'
'MAC' 'MAM' 'MAD' 'MAW' 'MAQ' 'MAP' 'MCG' 'MBZ' 'MBY' 'MAZ' 'MAX' 'MBX'	'CSS' 'CMA' 'CMX' 'CMC' 'CMV' 'EBF' 'DRS' 'DSM' 'DJS' 'ECO' 'EDS' 'EML'
'MCP' 'MDS' 'MEV' 'MEM' 'MCX' 'MFD' 'MFB' 'MEY' 'MFZ' 'MFV' 'MGA' 'MGS'	'EMM' 'EMC' 'EJM' 'EMF' 'EMA' 'EMS' 'EPS' 'ERG' 'ERV' 'FCC' 'FAR' 'EMX'
'MGL' 'MGR' 'MHF' 'MGX' 'MGB' 'MHG' 'MHM' 'MID' 'MHX' 'MIB' 'MJJ' 'MJV'	'FMM' 'HMB' 'HMT' 'GCP' 'FTX' 'HMF' 'HMC' 'FST' 'FTM' 'ISC' 'HPM' 'IMM'
'MJG' 'MIN' 'MJZ' 'MJY' 'MLP' 'MLB' 'MLR' 'MLX' 'MMK' 'MLZ' 'MLW' 'MMM'	'IBS' 'ISG' 'HRM' 'IMS' 'IRP' 'ISS' 'JMW' 'JFQ' 'JMC' 'JMS' 'JRS' 'LCP'
'MMY' 'MMW' 'MMR' 'MNM' 'MNA' 'MNC' 'AAW' 'ACL' 'ACP' 'AAV' 'ACX' 'AFY'	'LCC' 'LCR' 'LMF' 'LFC' 'LMR' 'MAK' 'LRS' 'LOS' 'LMX' 'MAF' 'MAL' 'LMY'
'AFX' 'AFC' 'AFD' 'AFA' 'ADS' 'AGR' 'AJR' 'AJT' 'ALS' 'AJJ' 'APC' 'AMU'	'LMV' 'MAA' 'MBD' 'MAY' 'MAO' 'MBN' 'MBF' 'MAR' 'MBS' 'MCJ' 'MCD' 'MCB'
'AMK' 'AMR' 'ANC' 'ATM' 'ATX' 'AXC' 'ATG' 'AWC' 'BCS' 'CAM' 'CAZ' 'CMM'	'MCF' 'MCA' 'MCU' 'MCT' 'MCS' 'MCR' 'MCV' 'MER' 'MCY' 'MDG' 'MES' 'MEG'
'CMP' 'CAR' 'CPN' 'COX' 'CSL' 'CRX' 'DAS' 'DEM' 'DFL' 'DFO' 'DFB' 'DPP'	'MEF' 'MEP' 'MCW' 'MEZ' 'MFF' 'MFC' 'MFG' 'MFA' 'MFS' 'MFR' 'MFP' 'MFO'
'FEM' 'FJM' 'ENS' 'FVC' 'FJQ' 'FJF' 'FSF' 'FOF' 'GTS' 'HMG' 'HML' 'JAA'	'MFJ' 'MFM' 'MFW' 'MGM' 'MFX' 'MGP' 'MIG' 'MHC' 'MGF' 'MGC' 'MHR' 'MIF'
'JJS' 'JJA' 'JAX' 'JFR' 'JMG' 'JGX' 'JDX' 'JJM' 'JMD' 'JMK' 'JPY' 'JSB'	'MHH' 'MIM' 'MIP' 'MIR' 'MJC' 'MJD' 'MJB' 'MIS' 'MJF'
'JSP' 'JMZ' 'JNS' 'JSG' 'JSV' 'JPX' 'JPC'	

Test set files' name (141 files):

Normal samples' file name (72 files)	Pathological samples' file name (69 files)
'MOJ' 'MOO' 'MNY' 'MPF' 'MRF' 'MSB' 'MRL' 'MRZ' 'MRY' 'MSM' 'MSP' 'MSU'	'JPF' 'JPM' 'JPS' 'JRT' 'JRP' 'JSC' 'JSS' 'LAS' 'JSW' 'LFS' 'LLR' 'MAS'
'MSG' 'MSL' 'MTC' 'MTB' 'MTO' 'MVA' 'MTR' 'MUR' 'MTG' 'MVN' 'MVM' 'MVF'	'MCL' 'MCZ' 'MJS' 'MMF' 'MMX' 'NHS' 'NVA' 'MVT' 'MJM' 'MJR' 'MKF' 'MJO'
'MXL' 'MWS' 'MXB' 'MWC' 'MXF' 'MWL' 'MXR' 'MYF' 'MYD' 'MYC' 'MXM' 'MZB'	'MJL' 'MLY' 'MLK' 'MLS' 'MLM' 'MMC' 'MMA' 'MMB' 'MNX' 'MMS' 'MMN' 'MNS'
'MYS' 'MZC' 'MYL' 'NMS' 'MZS' 'OMA' 'PAC' 'PBS' 'PCO' 'PCG' 'PFN' 'PDB'	'MNF' 'MMV' 'MOC' 'MOF' 'MPA' 'MOS' 'MOP' 'MOT' 'MPR' 'MPB' 'MPP' 'MRC'
'PGP' 'PIN' 'JSX' 'LBX' 'LAM' 'JTB' 'LMD' 'LBC' 'LNR' 'MFT' 'MFL' 'MJX'	'MRA' 'MRM' 'MPT' 'MRB' 'MRT' 'MRV' 'MSA' 'MRP' 'MRS' 'MRX' 'MSC' 'MSS'
'MJT' 'MLA' 'MMG' 'MPS' 'MPV' 'MPX' 'MSX' 'NMP' 'PFS' 'PMS' 'PCX' 'PDG'	'MSV' 'MSF' 'MSW' 'MSZ' 'MTF' 'MTA' 'MSY' 'MTS' 'MTM'

Validation set files' name (105 files):

Normal samples' file name (55 files)	Pathological samples' file name (50 files)
'SPS' 'PLS' 'PMC' 'RAS' 'POM' 'PMF' 'REP' 'RBS' 'RFM' 'RMG' 'ROM' 'RNG'	'ZSQ' 'PAF' 'OMN' 'PJG' 'PJR' 'PLB' 'RNA' 'RPA' 'RPM' 'SMF' 'SMR' 'ZPM'
'RPS' 'SAF' 'RVS' 'SBM' 'SIC' 'SIL' 'SCM' 'SCN' 'SIS' 'SRB' 'TJM' 'TLL'	'VQC' 'VMR' 'MVC' 'MVG' 'MXS' 'MXC' 'OFG' 'OAV' 'PCP' 'PCC' 'OMV' 'PAT'
'TJS' 'TIP' 'TRT' 'TSS' 'TSP' 'TMB' 'VDM' 'VCD' 'VAF' 'VFM' 'VIR' 'VLF'	'PJP' 'RAO' 'PMX' 'PSC' 'RAA' 'RFS' 'RDS' 'RCS' 'RCM' 'RFX' 'RLB' 'RMR'
'VMM' 'RFF' 'PSM' 'RAC' 'RRM' 'RBF' 'SEF' 'SMT' 'SMG' 'RTG' 'RSS' 'TDF'	'SAP' 'SAS' 'SCA' 'SAD' 'SMS' 'SPB' 'SMW' 'SMM' 'SRP' 'SRR' 'SRG' 'TIR'
'TDS' 'SSS' 'TRM' 'TMT' 'TAS' 'VPR' 'VMC'	'TFS' 'VLC'

Data availability The *AVFAD* dataset was distributed through the *ACSA* https://acsa.web.ua.pt/AVFAD.htm *platform.*

Declarations

Conflict of Interest The authors have no competing interests to declare that are relevant to the content of this paper.

References

Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. *IEEE Access, 10*, 122136–122158.

Abdulmajeed, N. Q., Al-Khateeb, B., & Mohammed, M. A. (2022). A review on voice pathology: Taxonomy, diagnosis, medical procedures and detection techniques, open challenges, limitations, and recommendations for future directions. *Journal of Intelligent Systems, 31*(1), 855–875.

Abdulmajeed, N. Q., Al-Khateeb, B., & Mohammed, M. A. (2023). Voice pathology identification system using a deep learning approach based on unique feature selection sets. *Expert Systems*. https://doi.org/10.1111/exsy.13327

Ali, Z., Alsulaiman, M., Muhammad, G., Elamvazuthi, I., & Mesallam, T. A. (2013). Vocal fold disorder detection based on continuous speech by using MFCC and GMM. In *2013 7th IEEE GCC conference and exhibition (GCC)*. IEEE.

Ali, Z., Elamvazuthi, I., Alsulaiman, M., & Muhammad, G. (2016). Automatic voice pathology detection with running speech by using estimation of auditory spectrum and cepstral coefficients based on the all-pole model. *Journal of Voice, 30*(6), 757.

Ali, Z., Hossain, M. S., Muhammad, G., & Sangaiah, A. K. (2018). An intelligent healthcare system for detection and classification to discriminate vocal fold disorders. *Future Generation Computer Systems, 85*, 19–28.

Ali, Z., Muhammad, G., & Alhamid, M. F. (2017). An automatic health monitoring system for patients suffering from voice complications in smart cities. *IEEE Access, 5*, 3900–3908.

Al-Sabaawi, A., Ibrahim, H. M., Arkah, Z. M., Al-Amidie, M., & Alzubaidi, L. (2020). Amended convolutional neural network with global average pooling for image classification. In *International conference on intelligent systems design and applications*. Springer.

Ankışhan, H., & İnam, S. Ç. (2021). Voice pathology detection by using the deep network architecture. *Applied Soft Computing, 106*, 107310.

Bengio, Y., Goodfellow, I., & Courville, A. (2017). *Deep learning*. MIT Press Cambridge.

Chen, J., Ye, J., Tang, F., & Zhou, J. (2021). *Automatic detection of Alzheimer's disease using spontaneous speech only*. NIH Public Access.

Chen, L., & Chen, J. (2022). Deep neural network for automatic classification of pathological voice signals. *Journal of Voice, 36*(2), 288.

Chuang, Z.-Y., Yu, X.-T., Chen, J.-Y., Hsu, Y.-T., Xu, Z.-Z., Wang, C.-T., Lin, F.-C., & Fang, S.-H. (2018). Dnn-based approach to detect and classify pathological voice. In *2018 IEEE international conference on big data (Big Data)*. IEEE.

Chugh, G., Kumar, S., & Singh, N. (2021). Survey on machine learning and deep learning applications in breast cancer diagnosis. *Cognitive Computation, 13*(6), 1451–1470.

Deepa, P., & Khilar, R. (2022). Speech technology in healthcare. *Measurement: Sensors, 24*, 1565.

Association, A. S.-L.-H. (2009). Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ASHA Special Interest Group 3, Voice and Voice Disorders. American Speech-Language-Hearing Association.

Graves, A., Fernández, S., & Schmidhuber, J. (2005). *Bidirectional LSTM networks for improved phoneme classification and recognition*. Springer.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks, 18*(5–6), 602–610.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., & Cai, J. (2018). Recent advances in convolutional neural networks. *Pattern Recognition, 77*, 354–377.

Hegde, S., Shetty, S., Rai, S., & Dodderi, T. (2019). A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice, 33*(6), 947.

Islam, R., Tarique, M., & Abdel-Raheem, E. (2020). A survey on signal processing based pathological voice detection techniques. *IEEE Access, 8*, 66749–66776.

Jesus, L. M., Barney, A., Santos, R., Caetano, J., Jorge, J., & Couto, P. S. (2009). Universidade de Aveiro's voice evaluation protocol. In *Tenth annual conference of the international speech communication association (Interspeech)*.

Jesus, L. M., Belo, I., Machado, J., & Hall, A. (2017). The advanced voice function assessment databases (AVFAD): Tools for voice clinicians and speech research. *Advances in Speech-Language Pathology*.

Kent, R. D. (2004). *The MIT encyclopedia of communication disorders*. MIT Press.

Ksibi, A., Hakami, N. A., Alturki, N., Asiri, M. M., Zakariah, M., & Ayadi, M. (2023). Voice pathology detection using a two-level classifier based on combined CNN–RNN architecture. *Sustainability, 15*(4), 3204.

Latif, S., Qadir, J., Qayyum, A., Usama, M., & Younis, S. (2020). Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering, 14*, 342–356.

Lee, J.-Y. (2021). Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the saarbruecken voice database. *Applied Sciences, 11*(15), 7149.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems, 33*(12), 6999–7019.

Mesallam, T. A., Farahat, M., Malki, K. H., Alsulaiman, M., Ali, Z., Al-Nasheri, A., & Muhammad, G. (2017). Development of the Arabic voice pathology database and its evaluation by using speech features and machine learning algorithms. *Journal of Healthcare Engineering, 2017*(1), 878351.

Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., Khanapi Abd Ghani, M., Maashi, M. S., Garcia-Zapirain, B., Oleagordia, I., Alhakami, H., & Al-Dhief, F. T. (2020). Voice pathology detection and classification using convolutional neural network model. *Applied Sciences, 10*(11), 3723.

Muhammad, G., & Alhussein, M. (2021). Convergence of artificial intelligence and internet of things in smart healthcare: A case study of voice pathology detection. *IEEE Access, 9*, 89198–89209.

Narendra, N., & Alku, P. (2020). Glottal source information for pathological voice detection. *IEEE Access, 8*, 67745–67755.

Oliveira, B. F., Magalhães, D. M., Ferreira, D. S., & Medeiros, F. N. (2020). Combined sustained vowels improve the performance of the Haar wavelet for pathological voice characterization. In *2020 International conference on systems, signals and image processing (IWSSIP)*, IEEE.

Payten, C. L., Chiapello, G., Weir, K. A., & Madill, C. J. (2022). Frameworks, terminology and definitions used for the classification of voice disorders: A scoping review. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2022.02.009

Ribas, D., Miguel, A., Ortega, A., & Lleida, E. (2023a). On the problem of data availability in automatic voice disorder detection. *In HEALTHINF,* (pp. 330–337).

Ribas, D., Pastor, M. A., Miguel, A., Martínez, D., Ortega, A., & Lleida, E. (2023b). Automatic voice disorder detection using self-supervised representations. *IEEE Access, 11*, 14915–14927.

Shekofteh, Y., & Almasganj, F. (2013). Remote diagnosis of unilateral vocal fold paralysis using matching pursuit based features extracted from telephony speech signal. *Scientia Iranica, 20*(6), 2051–2060.

Sindhu, I., & Sainin, M. S. (2024). Automatic speech and voice disorder detection using deep learning—a systematic literature review. *IEEE Access, 12*, 49667–49681.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.

Syed, S. A., Rashid, M., Hussain, S., & Zahid, H. (2021). Comparative analysis of CNN and RNN for voice pathology detection. *BioMed Research International, 2021*, 1–8.

Verma, V., Benjwal, A., Chhabra, A., Singh, S. K., Kumar, S., Gupta, B. B., Arya, V., & Chui, K. T. (2023). A novel hybrid model integrating MFCC and acoustic parameters for voice disorder detection. *Scientific Reports, 13*(1), 22719.

Zakariah, M., Ajmi Alotaibi, Y., Guo, Y., Tran-Trung, K., & Elahi, M. M. (2022). An analytical study of speech pathology detection based on MFCC and deep neural networks. *Computational and Mathematical Methods in Medicine, 2022*, 7814952.

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). *Dive into deep learning*. Cambridge University Press.

Zhao, D., Qiu, Z., Jiang, Y., Zhu, X., Zhang, X., & Tao, Z. (2024). A depthwise separable CNN-based interpretable feature extraction network for automatic pathological voice detection. *Biomedical Signal Processing and Control, 88*, 105624.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.